Program for the International Assessment of Adult Competencies (PIAAC) Proficiency Scores[1]

In each of the three domains assessed—literacy, numeracy, and problem solving in technology-rich environments (PS-TRE)—proficiency is considered as a continuum of ability involving the mastery of information-processing tasks of increasing complexity. The continuum is comprised of the all the test questions or "items" placed on a scale of 0-500 for each of these domains. The top of the scale, 500, represents the point at which the most complex task could be placed on the continuum.

The proficiency of adults who take the assessment is estimated based on their performance on the small sample of items they are administered. Adults who are able to correctly answer items at a particular point on the scale may be able to complete some other more difficult items (that is, those with a higher scale score) that they were not administered, but their probability of success decreases as the distance between their estimated average score and an item's placement on the scale increases. Conversely, adults will also have a greater chance to successfully complete easier items. Thus, the pool of PIAAC adults with an estimated average score of 300 will have one probability to give the right answer to an item with a score of 300, another (lower) probability to give the right answer to an item that is identified at the 400 score level, and another (higher) probability to give the right answer to an item that is identified at a 200 score level.

To report such performance results in a more meaningful way than scale score values, PIAAC results are reported in terms of *proficiency levels* (e.g., see Exhibit B-1 in the U.S. PIAAC report) that describe what adults can do at specific levels, in terms of the underlying skills needed at those levels. However, to say that adults performing at a certain estimated average score are able to successfully complete items at a specific proficiency level requires setting a 'cutoff' probability for reporting competency at that specific proficiency level. For example, one could use a response probability (RP) of 50 percent as a cutoff, but a 50 percent chance of success may not be deemed by all as demonstrating competency reliably. Contrarily, one could use a RP of 80 percent as a cutoff, but a 80 percent chance of success may be deemed by some as unfair because it classifies adults with a 75 percent chance of success as not reliably competent at that level.

This resource document provides an excerpt from the Technical Report of the Survey of Adult Skills (PIAAC) that explains how the cutoffs for PIAAC's proficiency levels were set and describes how they differ from those in other international studies.

[1] Taken from the Technical Report of the Survey of Adult Skills (PIAAC) [http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf], Organization for Economic Co-operation and Development (OECD) 2013.

21.2 Defining the proficiency levels

The item response theory (IRT) scaling procedures used in PIAAC constitute a statistical solution to the challenge of establishing a scale for a set of tasks with an order of difficulty that is essentially the same for everyone.

First, the response data collected from each participating country is used to estimate item parameters for each scale using a particular IRT model. In PIAAC, a two-parameter model is used, which models the probability of a response based on the difficulty of an item and how well it discriminates, in combination with the person's ability or proficiency. This information is summarized in the form of item characteristic curves, which show the probability of successfully completing an item at a given level of ability. Next, item parameters along with other information are used to estimate the ability distributions for each participating country along a scale with an overall mean and standard deviation. This scale can then be used to compare the overall performance of countries or subgroups within a country. It can also be used to compare performance along the scale based on statistical criteria such as percentiles.

The IRT analysis summarizes how well the sample of individuals who responded to the pool of tasks performed. The tasks in this pool constitute a sample of the universe or "population" of tasks representing the construct that is measured (in the case of PIAAC, literacy, numeracy and PS-TRE as defined by the relevant framework documents). Thus, the goal is to make inferences concerning the proficiency of respondents with respect to the population of tasks that represent the construct – that is, to make inferences about how well respondents performed on items used in the assessment as well as items having similar characteristics that also represent the construct but were not included in this particular assessment. As the items used in the survey represent a sample of tasks, it is important that the description of skills closely align to the framework used to define and construct them.

The use of IRT makes it possible not only to summarize results for various subpopulations of adults but also determine the relative difficulty of the tasks. In other words, just as individuals receive a specific score along a scale according to their performance on the assessment tasks, each task receives a specific value on a scale according to its difficulty, as determined by the performance of adults across the various countries that participated in the assessment (Kirsch et al., 2002). As tasks used in PIAAC vary widely in terms of task requirements and levels of complexity, it is possible to capture the range of difficulty of a task through an item map, which places all items along a scale based on a selected response probability.[2]

Test items do not discriminate perfectly and each person has a chance (however small) of responding correctly to any given item. Consequently, a value representing the probability of correctly responding to an item must be selected in order to place an item on a proficiency scale. In theory, any value greater than zero and less than one can be chosen to place items on a proficiency scale, and a range of response probability (RP)

[2] The RP section of this chapter was based on a PIAAC BPC document, Proficiency Levels in PIAAC [Doc. Ref.: COM/DELSA/EDU/PIAAC(2011)14], and written by Irwin Kirsch and Kentaro Yamamoto.

values are used in large-scale assessments[3]. A value of 0.62 is used in the Program for International Student Assessment (PISA) (OECD, 2009). Trends in International Mathematics and Science Study (TIMSS) uses different values for constructed responses (0.50) and multiple choice items (0.65) (TIMSS, 2007). The US National Assessment of Educational Progress (NAEP) uses an RP of 0.74 for multiple-choice items and 0.65 for open-ended items (National Center for Education Statistics, 2011). The International Adult Literacy Survey (IALS) and Adult Literacy and Lifeskills Survey (ALL) surveys used an RP of 0.80. The US National Assessment of Adult Literacy (NAAL) used an RP of 0.80 in reporting its 1992 survey and 0.67 in reporting results from its 2002 survey (Hauser, Edler, Koenig, & Elliott, 2005).

In PIAAC, the Organization for Economic Cooperation and Development (OECD) Secretariat and participating countries agreed on an RP value of 0.67, similar to the approach used in PISA, to ensure that the description of what it means to be performing at a particular level of proficiency is consistent between the two surveys. Given that both studies were developed and administered by the OECD, there were potential risks for the credibility of both studies if being at a particular level of proficiency meant something very different in each survey. While the RP value used in PIAAC and PISA are not identical[4], the interpretation of what it means to be at a level of proficiency is the same.

Within any given scale, except for those at the lowest level, a person would be expected to pass a test made up of items from the level at which he or she performed. For example, using RP67, a person at the bottom of Level 3 on the literacy scale would be expected to successfully complete items of Level 3 difficulty approximately 50 percent of the time, a person at the top of the level would be expected get such items correct around 80 percent of the time, and a person at the middle of the level would do so 67 percent of the time. In contrast, the probability of success on Level 3 items of persons at the top, bottom and middle of Level 3 based on RP80 is approximately 90, 60 and 80 percent, respectively. It is important to note that for both RP values, a person at the middle of a level would be likely to get most items at a lower level correct as well as a reasonable proportion of items at the next highest level correct. It is also important to note that the selection of a response probability is independent from the estimation of both item parameters and ability. The choice of an RP value has no impact on either the statistical characteristics of the items or the estimation of ability along the scale. In addition, the precision of measurement along a scale is not affected by the RP value. The same items define the underlying scale regardless of which RP value is selected.

As RP80 was used in IALS and ALL, in order to ensure that countries that wish to do so can map the change from RP80 to RP67, the OECD Secretariat provided item maps for literacy and numeracy under both the PIAAC approach (RP67) and the RP80 assumption (see Appendix A of The Survey of Adult Skills - Reader's Companion, pp. 107-12).

---

[3] Based on 'Literacy Levels and the 8- Percent Response Probability Convention [Doc. Ref.: http://nces.ed.gov/pubs2001/2001457_14.pdf], and written by Andrew Kolstad.

[4] This is a result of the different widths of the proficiency bands used.

References

Hauser, R. M., Edler, C. F. Jr., Koenig, J. A., & Elliott, S. W. (Eds.) (2005). *Measuring literacy: performance levels for adults*. Retrieved from http://www.nap.edu/catalog.php?record_id=11267

Kirsch I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change – performance and engagement across countries.* Retrieved from Organisation for Economic Co-operation and Development website: http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33690904.pdf

National Center for Education Statistics. (2011). *NAEP technical documentation*. Retrieved from NCES website: http://nces.ed.gov/nationsreportcard/tdw/analysis/describing_itemmapping.asp

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report.* Retrieved from OECD website: http://www.oecd.org/pisa/pisaproducts/pisa2006/42025182.pdf

Trends in International Mathematics and Science Study (TIMSS). (2007). TIMSS 2007 Technical report. Retrieved from TIMSS website: http://timss.bc.edu/timss2007/PDF/T07_TR_Chapter13.pdf